# Explaining and Upsampling Anomalies in Time-Series Sensor Data

Craig Pirie[1,2,*,†]

[1]Robert Gordon University, Garthdee House, Garthdee Road, Garthdee, Aberdeen, AB10 7AQ, Scotland, UK

**Abstract**

My research aims to improve anomaly detection methods in multi-sensor data by extending current re-sampling and explanation methods to work in a time-series setting. While there is a plethora of literature surrounding XAI for tabular data, the same cannot be said for the multivariate time-series settings. It is also known that selecting an optimal baseline for attribution methods such as integrated gradients remains an open research question. Accordingly, I am interested to explore the role of Case-Based Reasoning (CBR) in three ways: 1) to represent time series data from multiple sensors to enable effective anomaly detection; 2) to create explanation experiences (explanation-baseline pair) that can support the identification of suitable baselines to improve attribution discovery with integrated gradients for multivariate time-series settings; and 3) to represent the disagreements between past explanations in a case-base to better inform strategies for solving disagreement between explainers in the future. A common theme across my research is the need to explore how inherent relationships between sensors (causal or other ad-hoc inter-dependencies) can be captured and represented to improve anomaly detection and the follow-on explanation phases.

**Keywords**

anomaly detection, time-series, negative sampling, integrated gradients

## 1. Introduction

The advent of the Internet of Things (IoT) has empowered connected technology with new capabilities[1] [6]. One application of IoT is for the smart-monitoring of sensors to detect and predict failures. This is called anomaly detection and is an important problem across many domains. Its purpose is to identify patterns in data that lie outwith the realms of expectation [1]. It has many applications, including in fraud detection [10], medical analysis [3], industrial settings [17] and in sensor networks [8]. In the context of sensors, anomaly detection refers to the problem of identifying faulty or damaged sensors. This can rely on *univariate* (one sensor) or *multivariate* (many sensors) data. It is hypothesised that in real-world industrial settings, multivariate data is advantageous for anomaly detection. This is because it allows for the capturing of the hidden inter-dependencies between sensors which may improve the

*ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France*

[*] Corresponding author.

c.pirie11@rgu.ac.uk (C. Pirie)
https://www.linkedin.com/in/craig-pirie-aberdeen/ (C. Pirie)
0000-0002-6799-0497 (C. Pirie)

[1] IoT refers to a network consisting of sensory, communication, networking and information-processing technologies [6].

performance of anomaly detection systems. For example, a key multi-variate time-series dataset that will be used for this project includes the smart-buildings dataset[2]. It contains features such as **zone_air_cooling_setpoint** and **zone_air_temperature_sensor** that when evaluated on their own may be difficult to identify an anomaly. However, when considered in relation to one-another, the anomaly becomes more obvious (you would expect the two values not to be wildly dissimilar).

CBR has been used successfully with time-series problems [9] where there is a need to represent one or more data streams to enable decision support. Existing representation techniques for time-series data often make use of feature extraction (e.g.moving average) or feature transformation (e.g. DCT) methods. Common to these problems are the lack of human-annotated data or limited access to instances or a serious class imbalance as is expected with anomaly detection. By definition, anomalous instances are rare. Ergo, it is natural that the number of recorded normal instances far outweighs that of the abnormal class (see Figure 1). This can cause bias and be problematic for deep learning methods. Usually this leads to a pre-processing step to bring balance to the data prior to the learning phase. This can involve up-sampling the negative class at the cost of introducing artificial data into the set, or down-sampling the positive class at the cost of discarding valuable data. Class imbalance in time-series settings is particularly difficult due to the temporal connection between instances. The time dimension must be considered when sampling to allow the learning of time-series models. 'Negative Sampling' [16] is one method of up-sampling the anomalous class. It is based on the 'Concentration Phenomenon' and artificially inflates the feature space by applying a $\pm\delta$, to each of the features of samples in the normal space. However, data points are shuffled through time as the application of the $\delta$ is a random process and the time-dimension is not carefully considered. I wish to extend this method to work in the time-series setting. One approach I wish to take is to learn a sequence embedding of the feature space on which I will perform the re-sampling on prior to decoding. This way the temporal information is maintained and time-series methods may be applied.

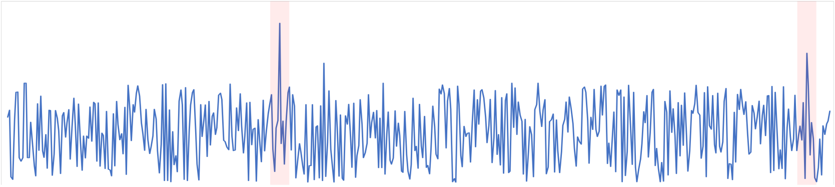

**Figure 1:** The figure shows a typical time-series representation in anomaly detection. Areas shaded in red are anomalies. This highlights the scarcity of abnormal instances compared to normal instances.

Before anomaly detection models can be deployed in production, we must learn to trust them. However, with the black-box nature of many deep neural architectures, this can be challenging. Confidence can be improved with Explainable AI (XAI). Through XAI, we can gain better insights around the 'decision-making process' of a deep-learning model. When we know why a decision is being made, we find it easier to trust that the decision was made on the correct grounds. Many strategies for explaining time-series re-purpose existing XAI methods (such as LIME [11] or SHAP [7]) that were originally used to explain models in other areas such as in

[2] The smart buildings dataset can be found here.

computer vision or natural language processing [12] and there is a lack of methods specifically designed for time-series. More so, they are often difficult to interpret and unsophisticated [14]. It has also been found that some saliency-based methods, that were designed for tabular data, fail when applied to sequence data [4, 2]. This can lead to anomalies being explained individually at a particular time-stamp, rather than collectively over a segment of a time-series graph. Increasingly, Case-based Reasoning (CBR) has seen application to facilitate XAI [15], giving rise to *XCBR*. Delaney *et al.* present a counterfactual-based XCBR approach to explaining time-series anomalies [2]. This works well for uni-variate data but interpretability suffers as the number of features increases. Several surveys have been conducted on the state-of-the-art in time-series explanation and there is a clear consensus that there is a need for more sophisticated methods that are dedicated to time-series applications [12, 13, 2].

Integrated gradients [19] is a popular XAI technique that does not modify the original network to provide explanations. It determines feature importance by evaluating the gradient between the input and output along uniform steps through the feature space. For this, it requires a baseline (often an all-zero embedding such as a black image). Alas, selection of a good baseline is problematic due to the *missingness* problem [18]. "Missingness" is a concept that is well-defined in game theory. It originally referred to the concept of determining how much value a group of participants added to a game by evaluating the value of the game after gradually adding more participants. This is the idea behind the baseline — to model the absence of the feature we are trying to evaluate. The problem arises when the same data resides in both the baseline *and* the sample because it becomes impossible to gradually introduce to determine its importance. In other words, $x_i - x_i'$ (the difference between baseline and output) will always be 0, meaning it can never be deemed important (see Figure 2 for an example). Therefore, it is essential that the baseline shares minimal information with the output image. There have been multiple strategies used to try accomplish this but none are perfect and the topic remains an open research area [18].

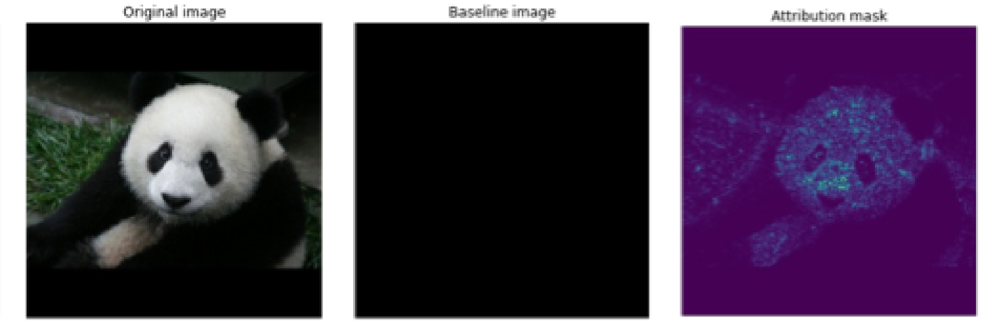

**Figure 2:** The original image (left) is of a giant panda and the baseline image (center) is a constant black image. The attribution mask (right) shows the pixels deemed important by integrated gradients by highlighting them. The figure demonstrates the missingness problem as the black features of the panda are deemed unimportant despite being a prominent characteristic.

To further ensure trust, often an ensemble approach is taken to explain models, meaning that explanations from multiple different explainers are used. This introduces the 'Disagreement Problem' [5] where it is common for many different explainers to produce wildly different explanations for the same decision. Ironically, this can further the distrust in systems — exactly the opposite of its purpose. There is a need for tackling this problem so that multiple explainers can be made to agree on an explanation. I will explore if this problem can be tackled by learning from past experiences, in a case-based manner. Through this, it is hoped that past experiences in solving explanation disputes can be used to solve disagreements in the future for similar problems. The case-base would consist of a model, its output and the different explanations as the query; and the strategy used to solve the disagreement as the solution.

Lastly, I wish to take a similar approach to learn the relational information in the feature space. Often, sensors in industrial domains do not work in solitude. There may be obvious, physical inter-dependencies between sensors or they may be hidden. For example, as humidity is proportionally affected by temperature, spikes in temperature readings not seen in humidity readings may indicate a broken temperature sensor (or vice-versa). By learning an embedding of the relational information I can capture these inter-dependencies when sampling in the encoded space. Consequently, it is hoped this additional information will support the classification ability of anomaly detection systems.

## 2. Research Plan

### 2.1. Research Objectives

The main aim of this project is to improve anomaly detection in industrial sensor data and its explanations by extending current methods to work with time-series data. Through this I will be examining how to represent sequence data (such as time-series) to enable re-sampling methods needed to improve coverage for anomaly detection. This will involve experiments with feature embeddings that capture temporal and relational information between sensors. I will also explore explainer aggregation strategies to address the Disagreement Problem in the time-series context. As such, the following research questions are defined:

- **RQ1:** How can sequence data (such as time-series) be represented to enable negative sampling methods that capture the temporal and relational information between sensors needed to improve anomaly detection?
- **RQ2:** Can local methods such as a case-based approach to selecting a baseline for Integrated Gradients improve the quality of its explanations?
- **RQ3:** What explainer aggregation strategies can be used to address the explainer Disagreement Problem in the context of time-series data for anomaly detection?

### 2.2. Approach / Methodology

I wish to conduct experiments with applying Negative Sampling on an embedding of the feature space that captures the temporal information and inter-dependencies between sensors. This may involve the use of auto-encoders or sequence-to-sequence auto-encoders to learn the embedding. It is hoped that incorporating this data will improve the performance anomaly classifiers.

Currently I propose to expand on Integrated Gradients for the time-series setting by deploying a case-based approach to selecting baselines. To do this, a query (time-series window, image etc.) and baseline pair will be stored in a case base. Similar images contain similar information, therefore, in theory, should react well to similar baselines. To use an image-based example, horses are very similar animals to zebras, so it is hoped that a baseline that provides good explanations for horses, can do so for zebras. For time-series, similar trends or sub-sequences will use similar baselines. This should circumvent the 'missingness problem' but may require an adaptation step to further ensure the missingness characteristic is upheld.

Finally, I will also explore the efficacy of case-based reasoning to solve the 'Disagreement Problem' in ensemble explainers. The aim is to establish a case-base of 'disagreements' and solutions (past strategies that were used to solve the disagreement).

## 3. Progress Summary

The research is still in its embryonic stages and we are still tweaking the research proposal. We have identified some key papers which we are reviewing in depth to better understand the problem and validate the need for our research. Accompanying this, some exploratory data analysis is being conducted in parallel with the hopes to replicate the results in studies found in the literature. At present, the main focus is evaluating negative sampling as a method to correct class imbalance and determining its suitability for time-series data.

## References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.

[2] Eoin Delaney, Derek Greene, and Mark T Keane. "Instance-based counterfactual explanations for time series classification". In: *International Conference on Case-Based Reasoning*. Springer. 2021, pp. 32–47.

[3] Milos Hauskrecht et al. "Evidence-based anomaly detection in clinical domains". In: *AMIA Annual Symposium Proceedings*. Vol. 2007. American Medical Informatics Association. 2007, p. 319.

[4] Aya Abdelsalam Ismail et al. "Benchmarking deep learning interpretability in time series predictions". In: *Advances in neural information processing systems* 33 (2020), pp. 6441–6452.

[5] Satyapriya Krishna et al. "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective". In: *arXiv preprint arXiv:2202.01602* (2022).

[6] Shancang Li, Li Da Xu, and Shanshan Zhao. "The internet of things: a survey". In: *Information systems frontiers* 17.2 (2015), pp. 243–259.

[7] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[8] Luis Martí et al. "Anomaly detection based on sensor data in petroleum industry applications". In: *Sensors* 15.2 (2015), pp. 2774–2797.

[9] Stewart Massie et al. "Monitoring Health in Smart Homes using Simple Sensors". In: ().

[10] Tahereh Pourhabibi et al. "Fraud detection: A systematic literature review of graph-based anomaly detection approaches". In: *Decision Support Systems* 133 (2020), p. 113303.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[12] Thomas Rojat et al. "Explainable artificial intelligence (xai) on timeseries data: A survey". In: *arXiv preprint arXiv:2104.00950* (2021).

[13] Udo Schlegel et al. "An empirical study of explainable AI techniques on deep learning models for time series tasks". In: *arXiv preprint arXiv:2012.04344* (2020).

[14] Udo Schlegel et al. "Towards A Rigorous Evaluation Of XAI Methods On Time Series". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019, pp. 4197–4201. DOI: 10.1109/ICCVW.2019.00516.

[15] Jakob M Schoenborn et al. "Explainable case-based reasoning: a survey". In: *AAAI-21 Workshop Proceedings*. 2021.

[16] John Sipple. "Interpretable, multidimensional, multimodal anomaly detection with negative sampling for detection of device failure". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9016–9025.

[17] Ljiljana Stojanovic et al. "Big-data-driven anomaly detection in industry (4.0): An approach and a case study". In: *2016 IEEE international conference on big data (big data)*. IEEE. 2016, pp. 1647–1652.

[18] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. "Visualizing the Impact of Feature Attribution Baselines". In: *Distill* (2020). https://distill.pub/2020/attribution-baselines. DOI: 10.23915/distill.00022.

[19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *CoRR* abs/1703.01365 (2017). arXiv: 1703.01365. URL: http://arxiv.org/abs/1703.01365.